Available online at www.sciencedirect.com

ScienceDirect

Applied Mathematics Letters 20 (2007) 312–315

Applied Mathematics Letters

www.elsevier.com/locate/aml

An improved class of estimators of a finite population quantile in sample surveys☆

A. Arcos*, M. Rueda, J.F. Muñoz

Department of Statistics and Operational Research, University of Granada, Spain

Received 28 July 2005; received in revised form 19 April 2006; accepted 24 April 2006

Abstract

This work proposes a general class of estimators for a finite population quantile using auxiliary information. This information is provided by the population means of auxiliary variables. The optimum estimator in this class is derived. This result is supported with a numerical example.

Keywords: Auxiliary information; Difference type estimator; Finite population quantile; Horvitz–Thompson estimator; Sample surveys

1. Introduction

The utility of estimating population quantiles to describe the distribution of a population characteristic of interest is well known.

Most research into quantile estimation [1,5] deals exclusively with the interest variable and does not make explicit use of auxiliary variables in the construction of estimators. In practice, it is normal to have auxiliary information on one or more population characteristics related to the principal variable. In many situations, the information is associated with the interest variable but studied on a previous occasion when a census was carried out. Kuk and Mak [2], Rao et al. [3], Rueda et al. [4] and Singh et al. [7] have considered the problem of median estimation when the population median of an auxiliary variable is known.

In this work we suggest a class of estimators for a finite population quantile of the interest variable when the population means of the auxiliary variables are known.

2. A proposed class of estimators for the quantiles

Assume that a sample s of size n is drawn from a finite population U of size N by a specific sampling design d with first and second order positive inclusion probabilities, π_i and π_{ij}. Let y be the interest variable (which is the object of study) and x the auxiliary variable. For each $t \in \mathbb{R}$, the finite population distribution function of y is defined

☆ Research partially supported by CICE (Junta de Andalucía) contract no. SEJ565.
* Corresponding author. Tel.: +34 58243267; fax: +34 58249046.
E-mail address: arcos@ugr.es (A. Arcos).

0893-9659/$ - see front matter
doi:10.1016/j.aml.2006.04.014

by $F_y(t) = N^{-1}\sum_{i\in U}\Delta(t-y_i)$, with $\Delta(a)=1$ if $a\geq 0$ and $\Delta(a)=0$ otherwise. The finite population β-quantile, $Q_y(\beta)$ $(0<\beta<1)$, is defined by $Q_y(\beta)=F_y^{-1}(\beta)=\inf\{y_i\in U: F_y(y_i)\geq\beta\}$.

First, assume that there are $x_1,\ldots,x_k$ auxiliary variables, and that the finite population β-quantiles $Q_{xi}(\beta)$, $i=1,\ldots,k$ are unknown, but that the finite population means $\overline{X}_1,\ldots,\overline{X}_k$ of the auxiliary variables are known. The usual unbiased estimators (the Horvitz–Thompson estimators) of $\overline{X}_i$ and $Q_y(\beta)$ are $\overline{x}_{HTi}=N^{-1}\sum_{j\in s}x_{ij}/\pi_j$ and $\widehat{Q}_y(\beta)=\widehat{F}^{-1}_{HTy}(\beta)=\inf\{y_i\in s:\widehat{F}_{HTy}(y_i)\geq\beta\}$, respectively, with $\widehat{F}_{HTy}(t)=N^{-1}\sum_{j\in s}\Delta(t-y_j)/\pi_j$.

Following the approach adopted by Srivastava and Jhajj [8], we propose a family of estimators of $Q_y(\beta)$ as

$$t_w = H(\widehat{Q}_y(\beta), u_1,\ldots,u_k), \tag{1}$$

where $u_i=\overline{x}_{HTi}/\overline{X}_i$, and $H(\cdot)$ is a function that is continuous and bounded in a closed subset $P\subset\mathbb{R}^{k+1}$ containing the point $(Q_y(\beta),\overline{X}_1,\ldots,\overline{X}_k)$ such that:

- $H(Q_y(\beta),1,\ldots,1)=Q_y(\beta)$ and the first order partial derivative of H with respect to $\widehat{Q}_y(\beta)$, $H_0(Q_y(\beta),1,\ldots,1)=1$.
- The first and second order partial derivatives of $H(\widehat{Q}_y(\beta),u_1,\ldots,u_k)$ exist and are continuous and bounded in P.

Note that $\widehat{Q}_y(\beta)$ and $\widehat{Q}_{ri}(\beta)=\widehat{Q}_y(\beta)\overline{X}_i/\overline{x}_{HTi}$, $\forall i$, are included in (1).

Expanding this function H about the point $(Q_y(\beta),1,\ldots,1)$ in a second order Taylor series we obtain that the bias of t_w is of order $O(n^{-1})$ and the mean squared error of t_w, to the first degree of approximation, is given by

$$\mathrm{MSE}(t_w)=V(\widehat{Q}_y(\beta))+H'\mathbf{C}H+2H'\mathbf{C}_0,$$

where $H'=(H_1,\ldots,H_k)$, H_i is the first order partial derivative of H with respect to u_i at the point $(Q_y(\beta),1,\ldots,1)$, $\mathbf{C}=(c_{ij})_{k\times k}$, $\mathbf{C}_0=(c_{01},\ldots,c_{0k})'$, $c_{0i}=\mathrm{cov}\left(\widehat{Q}_y(\beta),\overline{x}_{HTi}\right)Q_y(\beta)/\overline{X}_i$ and $c_{ij}=\mathrm{cov}\left(\overline{x}_{HTi},\overline{x}_{HTj}\right)Q_y(\beta)^2/(\overline{X}_i\overline{X}_j)$, $i=1,\ldots,k$.

Proposition 1. *Up to terms of order* $O(n^{-1})$,

$$\mathrm{MSE}(t_w)\geq V(\widehat{Q}_y(\beta))-\mathbf{C}_0'\mathbf{C}^{-1}\mathbf{C}_0. \tag{2}$$

Proof. The optimum value of H can be obtained by differentiating the above expression for the $\mathrm{MSE}(t_w)$ and equating to zero. We obtain $H=-\mathbf{C}^{-1}\mathbf{C}_0$. □

Exponential and difference type estimators $\widehat{Q}_{y_{ex1}}(\beta)=\widehat{Q}_y(\beta)\prod_{i=1}^k\left(\overline{X}_i/\overline{x}_{HTi}\right)^{\alpha_i}$ and $\widehat{Q}_{y_{dif1}}(\beta)=\widehat{Q}_y(\beta)+\sum_{i=1}^k c_i(\overline{x}_{HTi}-\overline{X}_i)$, are also included in (1). The optimum values of α_i and c_i $(i=1,\ldots,k)$ can be obtained by minimizing $\mathrm{MSE}(t_w)$ with a respective function H. The resulting estimators with the optimum constants have the minimum MSE, which is given by (2). Thus, asymptotically, $\widehat{Q}_{y_{ex1}}(\beta)$ and $\widehat{Q}_{y_{dif1}}(\beta)$ are optimum in this class.

3. The proposed estimators under simple random sampling

Under SRSWOR design, $\overline{x}_{HTi}$ reduces to the sample mean, $\overline{x}_i=n^{-1}\sum_{j\in s}x_{ij}$, $\widehat{F}_{HTy}(t)$ reduces to the ordinary sample empirical distribution function, and $\widehat{Q}_y(\beta)=\widehat{F}_y^{-1}(\beta)$ reduces to the sample β-quantile of y. In this particular case, variances and covariances are easily obtained (see the Appendix):

$$v(\overline{x}_i)=\frac{1-f}{n}S_{xi}^2,\qquad v(\widehat{Q}_y(\beta))=\frac{1-f}{n}\beta(1-\beta)\left(\frac{1}{f_y(Q_y(\beta))}\right)^2,$$

$$\mathrm{cov}(\overline{x}_i,\overline{x}_j)=\frac{1-f}{n}S_{xixj},\qquad \mathrm{cov}(\overline{x}_i,\widehat{Q}_y(\beta))=-\frac{1-f}{n}\frac{\rho_{zxi}S_{xi}\sqrt{\beta(1-\beta)}}{f_y(Q_y(\beta))},$$

where $z_j=\Delta(Q_y(\beta)-y_j)$, ρ_{zxi} denotes the coefficient of correlation between z and x_i. Variances and covariances can be estimated from the sample and $f_y(Q_y(\beta))$ can be approximated following Silverman [6].

Asymptotic properties of estimators in the proposed class are derived assuming that the finite population embeds in a sequence of populations $\{U_\nu\}$, where n_ν and N_ν increase such that $\frac{n_\nu}{N_\nu}\to f$ when $n_\nu\to\infty$. It is also assumed

that when $N_v \to \infty$, the multivariate distribution can be approximated by a continuous distribution with marginal densities f_y and f_{x_i} for y and x_i, $(i = 1, \ldots, k)$ respectively, and $f_{x_i}(Q_{x_i}(\beta))$ are positive.

Under simple random sampling, the estimators in the proposed class t_w are asymptotically unbiased and normal. First, the asymptotic unbiasedness of the proposed class of estimators is easily derived from its linear expression and since the estimators $\widehat{Q}_y(\beta)$ and $\overline{x}_i$ are, respectively, asymptotically unbiased and unbiased for their respective parameters. Second, as $\overline{x}_i$ and $\widehat{Q}_y(\beta)$ are asymptotically normal [1], then so are the estimators in t_w.

4. A more general class of estimators

We consider the situation in which the population means and the population β-quantiles associated with the auxiliary variables x_i, $\overline{X}_i$ and $Q_{xi}(\beta)$ are known. A family of estimators of $Q_y(\beta)$ that is broader than t_w is defined by

$$T_w = H(\widehat{Q}_y(\beta), u_1, \ldots, u_k, v_1, \ldots, v_k), \tag{3}$$

where $u_i = \overline{x}_{HTi}/\overline{X}_i$, $v_i = \widehat{Q}_{xi}(\beta)/Q_{xi}(\beta)$ and $H(\widehat{Q}_y(\beta), u_1, \ldots, u_k, v_1, \ldots, v_k)$ is a function that satisfies similar regularity conditions to those given in Section 3. Eq. (3) provides a class of estimators, of which t_w is a particular case. This class also includes:

- the ratio estimator proposed by Kuk and Mak [2], $\hat{Q}_r(\beta) = \hat{Q}_y(\beta)\frac{Q_{xi}(\beta)}{\hat{Q}_{xi}(\beta)}$,
- the difference estimator proposed by Rao et al. [3],

$$\hat{Q}_{dr}(\beta) = \hat{Q}_y(\beta) + \hat{R}\left(Q_x(\beta) - \hat{Q}_x(\beta)\right), \quad \text{with } \hat{R} = \frac{\sum_{i\in s} y_i/\pi_i}{\sum_{i\in s} x_i/\pi_i},$$

- the regression estimator proposed by Rueda et al. [4],

$$\hat{Q}_{\text{reg}}(\beta) = \hat{Q}_y(\beta) + \hat{b}\left(Q_x(\beta) - \hat{Q}_x(\beta)\right), \quad \text{with } \hat{b} = \frac{\sum_{i\in s} x_i y_i/\pi_i}{\sum_{i\in s} x_i^2/\pi_i}.$$

We can write (3) as $T_w = H(\widehat{Q}_y(\beta), w_1, \ldots, w_{2k})$ with $w_i = u_i$ for $i = 1, \ldots, k$ and $w_i = v_i$ for $i = k+1, \ldots, 2k$. The bias and the MSE of the estimators in the class T_w can be derived using a method similar to the one followed in the previous section.

Exponential and difference estimators:

$$\widehat{Q}_{y_{ex2}}(\beta) = \widehat{Q}_y(\beta)\prod_{i=1}^{k}\left(\frac{\overline{X}_i}{\overline{x}_{HTi}}\right)^{\gamma_i}\prod_{i=1}^{k}\left(\frac{Q_{x_i}(\beta)}{\widehat{Q}_{x_i}(\beta)}\right)^{\delta_i},$$

$$\widehat{Q}_{y_{dif2}}(\beta) = \widehat{Q}_y(\beta) + \sum_{i=1}^{k} a_i(\widehat{Q}_{x_i}(\beta) - Q_{x_i}(\beta)) + \sum_{i=1}^{k} d_i(\overline{x}_{HTi} - \overline{X}_i),$$

are also asymptotically optimum in T_w. Thus, asymptotically, $\widehat{Q}_{y_{ex2}}(\beta)$ and $\widehat{Q}_{y_{dif2}}(\beta)$ are more efficient than all the cited estimators.

5. Numerical comparisons

The following examples reflect the potential gains from the use of the proposed estimators instead of the customary estimators. Three well known populations are considered (see [4] for details): SUGAR CANE, MU281 and FAM1500.

Let consider us SRSWOR and the median estimation, $Q_y(0.5) = M_y$. Assuming that the mean or the mean and the median associated with an auxiliary variable, $\overline{X}$ and M_x, are known, the following five estimators are compared: the simple estimator, $\widehat{M}_y$, the usual difference ($\widehat{M}_d$) and ratio ($\widehat{M}_r$) estimators and the proposed estimators $\widehat{M}_{D1} = \widehat{M}_y + \widehat{c}_{opt}\left(\overline{X} - \overline{x}\right)$, and $\widehat{M}_{D2} = \widehat{M}_y + \widehat{a}_{opt}\left(M_x - \widehat{M}_x\right) + \widehat{d}_{opt}\left(\overline{X} - \overline{x}\right)$. The root of the ratio, R, of the variance to the variance of the simple estimator is computed as a measure of efficiency, $\text{R}(\widehat{\theta}) = \sqrt{V(\widehat{\theta})/V(\widehat{M}_y)}$.

Table 1
Relative efficiencies (R) of the various estimators when the median is estimated

	$\widehat{M}_y$	$\widehat{M}_d$	$\widehat{M}_r$	$\widehat{M}_{D1}$	$\widehat{M}_{D2}$
SUGAR CANE	1.0000	0.7882	1.1848	0.8058	0.7525
MU281	1.0000	0.9125	0.9317	0.8781	0.8766
FAM1500	1.0000	0.7477	0.8023	0.7074	0.6912

From Table 1 we deduce that the proposed estimators are more efficient than the simple estimator which does not use information of the population means. Evidently, $\widehat{M}_{D2}$ is the most accurate since this estimator uses population information (mean and median) from the auxiliary variable.

Appendix

In this appendix, $\operatorname{cov}(\overline{x}_i, \widehat{Q}_y(\beta))$ is derived. The Taylor series expansion yields

$$\widehat{Q}_y(\beta) - Q_y(\beta) \simeq \frac{1}{f_y(Q_y(\beta))}\left(\widehat{F}_y(\widehat{Q}_y(\beta)) - \widehat{F}_y(Q_y(\beta))\right) + O(n^{-\frac{1}{2}}) \simeq \frac{\left(\beta - \overline{z}_y\right)}{f_y(Q_y(\beta))}$$

where $\overline{z}_y = \widehat{F}_y(Q_y(\beta)) = n^{-1}\sum_{j\in s} z_j$ and $f_y(\cdot)$ is the derivative of $\widehat{F}_y(\cdot)$, the limiting value of $F_y(\cdot)$ as $N \to \infty$. Then

$$\begin{aligned} E\left(\overline{x}_i - \overline{X}_i\right)\left(\widehat{Q}_y(\beta) - Q_y(\beta)\right) &\simeq \frac{1}{f_y(Q_y(\beta))} E\left(\overline{x}_i - \overline{X}_i\right)\left(\beta - \overline{z}_y\right) \\ &= -\frac{\operatorname{cov}\left(\overline{x}_i, \overline{z}_y\right)}{f_y(Q_y(\beta))} = -\frac{1-f}{n}\frac{\rho_{zxi} S_{xi} S_z}{f_y(Q_y(\beta))} = -\frac{1-f}{n}\frac{\rho_{zxi} S_{xi}\sqrt{\beta(1-\beta)}}{f_y(Q_y(\beta))}. \end{aligned}$$

References

[1] S.T. Gross, Median estimation in sample survey, Proc. Surv. Res. Meth. Sect. Amer. Statist. Ass. (1980) 181–184.
[2] A. Kuk, T.K. Mak, Median estimation in the presence of auxiliary information, J. Roy. Statist. Soc. Ser. B 1 (1989) 261–269.
[3] J.N.K. Rao, J.G. Kovar, H.J. Mantel, On estimating distribution functions and quantiles from survey data using auxiliary information, Biometrika 77 (1990) 365–375.
[4] M. Rueda, A. Arcos, M.D. Martínez, Difference estimators of quantiles in finite populations, Test 12 (2003) 481–496.
[5] J. Sedransk, P.J. Smith, Inference for finite population quantiles, in: P.R. Krishnaiah, C.R. Rao (Eds.), Handbook of Statistics, vol. 6, North-Holland, 1988, pp. 267–289 (Chapter 11).
[6] B.W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman and Hall, London, 1986.
[7] S. Singh, A.H. Joarder, D.S. Tracy, Median estimation using double sampling, Aust. N. Z. J. Stat. 43 (2001) 33–46.
[8] S.K. Srivastava, H.S. Jhajj, A class of estimators of the population mean in survey sampling using auxiliary information, Biometrika 68 (1981) 341–343.